Dear authors,

The revised manuscript has adequately addressed a majority of the concerns from two reviewers during the first round of review. However, as you can see, there are still some issues which need to be further addressed, especially the specific comments from Anonymous referee #1. Hence, this MS is still subject to revisions before considering for publication.

Best regards,
Hongkai Gao

[Answer]
Dear Dr. Hongkai Gao,

We really appreciate your time handling our manuscript. Based on the Reviewer #1 comments, we carefully revised the manuscript including additional results of SWE comparison between Noah-MP runs and in-situ observations at a snow pillow site (Hubbard Brook, New Hampshire. We also re-ran the AMSR2 DA simulations to correct errors and revised Figures 4 & 7 with additional descriptions. For further details, please see our responses to relevant comments given by Reviewer #1.

We consider this modified version of the manuscript to be significantly improved because of those comments, and therefore hope that this revision will sufficiently address the reviewer's concerns. Thank you again for your time and efforts.

Sincerely,
Eunsang Cho, Yonghwan Kwon, Sujay Kumar, and Carrie Vuyovich

Second Review of: Assimilation of airborne gamma observations provides utility for snow estimation in forested environments

Overview: The revised manuscript has adequately addressed a majority of my major concerns during the first round of review, including performing some simple sensitivity analysis with Noah-MP to address the poor performance of the open loop simulation, and including some additional details regarding model set up and methods as well as significant steps to improve replicability of the study. Furthermore, the DA of gamma flightline measured SWE in forested regions of the Northeast US is of potential value since this can address a number of issues in the region related to snow characterization on fine scales. In particular, the result quantifying the impacts of localization distance on model performance can inform future data collection strategies and constrain regional SWE estimates from blended model/observational approaches. The DA approach is reasonable, and the gamma-SWE dataset is well validated and widely accepted within the community. Taken altogether, this study is of potential high-value to the community and worthy of publication. However, there are still some lingering larger concerns with the study that should be addressed prior to publication. Additionally, the modifications during the first round of revisions introduced a number of minor technical issues that need to be corrected.

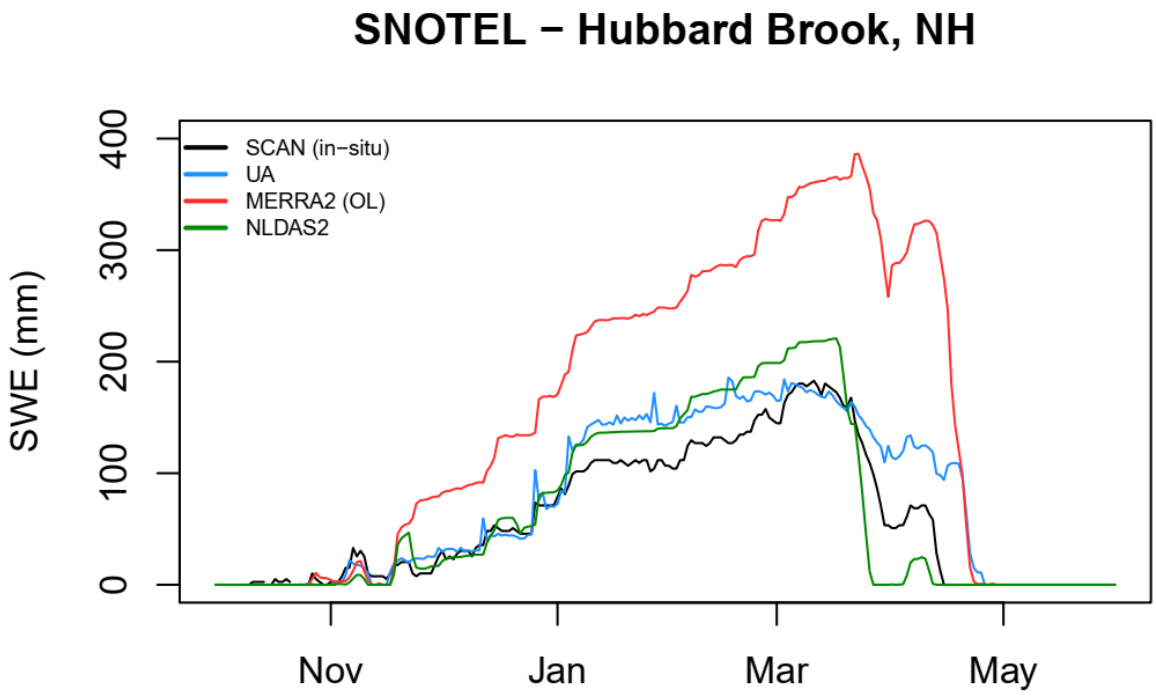
[Answer] We very much appreciate the Reviewer's time and valuable comments. We included additional results with in-situ SWE observations and another run with the NLDAS2 forcing data. We also addressed the reasons for showing some weird patterns in the SWE time series and re-ran the AMSR2 DA simulations with unit corrections. We hope that this revision will sufficiently address the Reviewer's concerns. Please see the details below.

Major comments:

1. While the authors have made substantial strides towards addressing the poor performance of the OL simulations, it is still concerning just how bad the model performance appears to be. While there are and can be fairly large model errors with Noah-MP, particularly around SWE max and during the melt season these are by far the most egregious that I can recall seeing in the literature. Accordingly, because the results of the OL simulation are so questionable for a widely used and accepted numerical model, I think the bar to publishing these results should be quite high. In the first revision, the authors have made solid efforts to add context to the model performance, however, to reach this bar, in addition to what the authors have already done, I recommend adding (or at least investigating, even if these results don't end up in the final manuscript) three specific things:

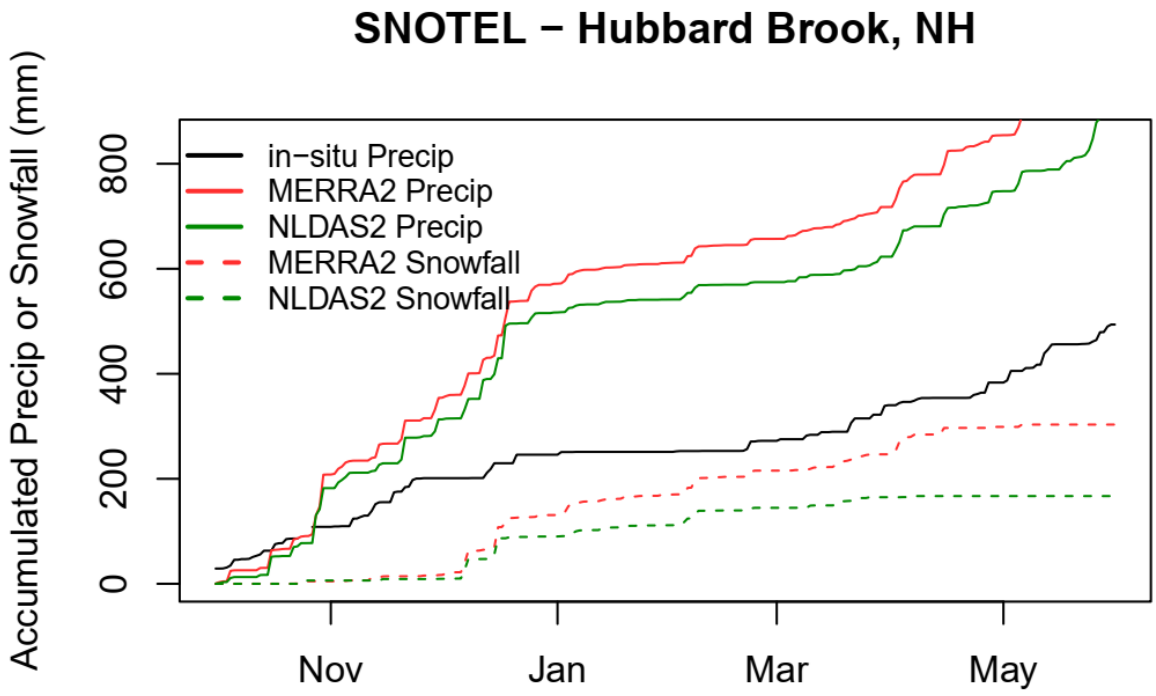
a. Find a single example location within the model domain where with an in-situ SWE measurement and compare SWE from the UA dataset, the OL simulation, and different model based product for snow (e.g., NLDAS). Snow depth could be used as a backup if there are no in-situ SWE measurements in the model domain.

[Answer] According to the Reviewer's suggestion, we found a SNOTEL/SCAN site at Hubbard Brook, NH where in-situ snow pillow SWE measurements are available, and made an SWE time series plot from Oct 1, 2002, to May 31, 2003, along with the UA observations, OL (MERRA2), and a new Noah-MP simulation forced by NLDAS2 forcing data. As we found previously, the OL-MERRA2 simulation overestimated SWE while the SWE simulations from NLDAS2 with the same Noah-MP parameterization options were close to the in-situ and UA SWE, which supports our previous analysis (in the first round of revision) that meteorological forcing is the main reason for the large overestimation of SWE during the accumulation period. However, the pattern of the rapid snow melting is also seen in the model results driven by NLDAS2.



b. Perform a simple "reality check" analysis of the MERRA-2 forcing against a handful (or even a single) in-situ weather station to explore possible biases associated with the reanalysis forcing in a more direct way than comparing LSM results from different versions.

[Answer] Thank you for the suggestion. To address this, we compared the MERRA2 forcing precipitation to the in-situ precipitation along with the NLDAS2 forcing precipitation below. The figure showed that both reanalysis forcings substantially overestimated precipitation between November and December. This overestimated precipitation led to overestimated SWE simulations particularly with MERRA-2.



c. Include one more Noah-MP simulation with the rain/snow partitioning threshold set to 0.0 instead of using the BATS = 2.5C threshold. I suspect that the lack of sensitivity to the precipitation phase partitioning is tied to the fact that both Jordan (1991) and BATS have that 2.5 threshold for rain/snow within them.

[Answer] To address the Reviewer's concern, we ran two more Noah-MP simulations with the rain/snow partitioning threshold set to 0.0 C with MERRA2 and NLDAS2, respectively. The figure below shows the SWE differences between Jordan 1991 (2.5 C) and 0.0 C. Based on the results, a change in the precipitation phase partitioning threshold from Jordan 1991 (2.5 C) to 0.0 C led to the SWE decrease up to 65 mm for MERRA2 and 20 mm for NLDAS2. This also confirms that the MERRA-2 forcing is responsible for the large overestimation of SWE rather than the parameterization option, which although contributes to a small portion of the SWE overestimation.

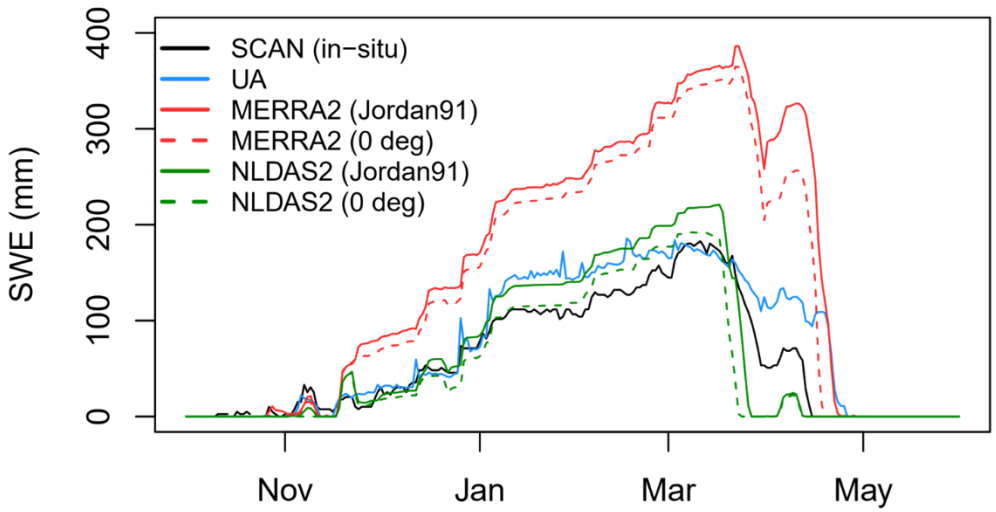


Figure 9. Comparison of SWE time series between four Noah-MP simulations and the Soil Climate Analysis Network (SCAN) ground-based observations at Hubbard Brook, New Hampshire from Oct 1, 2002, to May 31, 2003 (https://wcc.sc.egov.usda.gov/nwcc/site?sitenum=2069). The four Noah-MP SWE simulations were generated with Jordan (1991)'s scheme and a single threshold of 0°C and two meteorological forcings (MERRA2 – which is used for OL and the North American Land Data Assimilation System; NLDAS2), respectively.

Addressing these three items would do the following: 1) directly contextualize and ground truth the OL model results and the UA dataset with an on-the-ground SWE observation within the region, 2) illustrate, directly, possible biases in the model forcing, and 3) provide a full sensitivity analysis to phase-partitioning. Once these have been done, I think that frees up the authors to be more speculative regarding the model performance, and perhaps even make more general statements on how to improve the model in this region.

[Answer] Again, thank you for the valuable suggestions. Through these items, we found that the overestimated SWE simulations dominantly resulted from the reanalysis of meteorological forcings (precipitation) rather than the Noah-MP model itself.

We include the last figure in the revised manuscript (Figure 9) with descriptions as below.

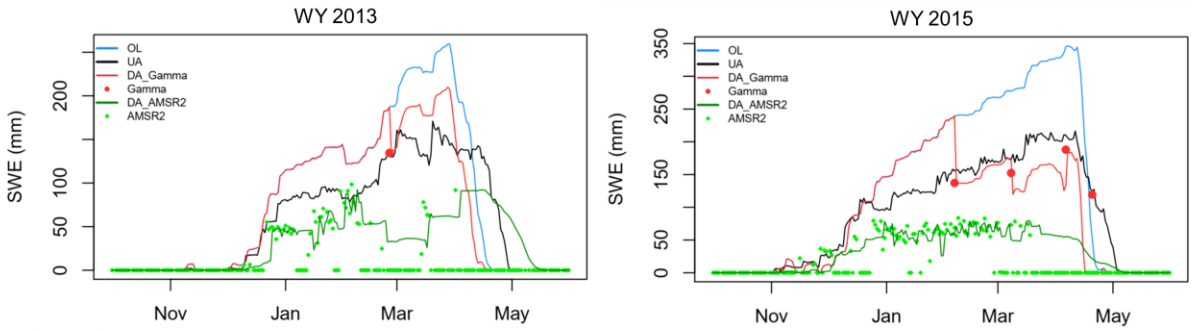
"Letcher et al. (2021) demonstrated that the use of cooler Tair thresholds in Noah-MP can improve the estimates of peak SWE in the northeastern United States. To verify this, four Noah-MP SWE simulations with Jordan (1991)'s scheme and a single threshold of 0°C with two different meteorological forcings (MERRA2 and the North American Land Data Assimilation System; NLDAS2) are compared to ground-based SWE observations from Oct 1, 2002, to May 31, 2003, at Hubbard Brook, New Hampshire, which is within the study domain (Figure 9). This supports the previous finding that the overestimated SWE with Jordan's scheme was reduced with a single threshold of 0°C for both forcings. This also presents that the use of regionally reliable meteorological forcings (e.g., precipitation) generates accurate SWE estimations.

2. To me there appears to be some weirdness going on in figure 4 that should be explained.

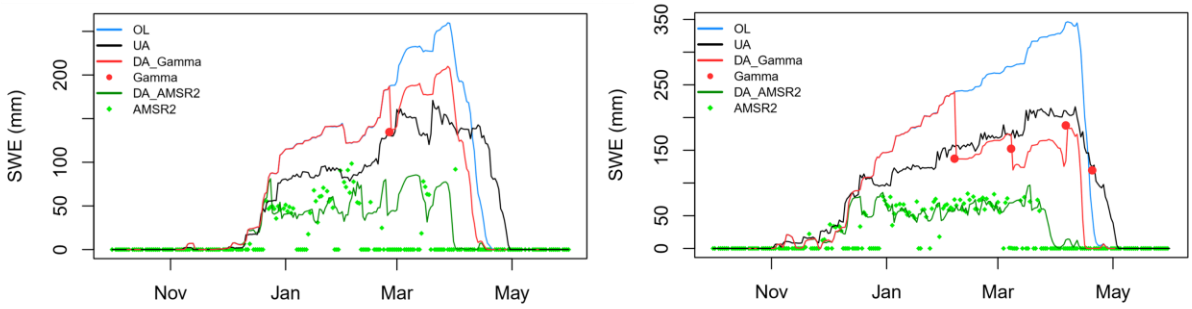
a. How does the AMSR-E DA run end up with a later melt out date than all other simulations despite consistently lower SWE throughout the season, and a number of "zero" SWE assimilations in the spring? That is, why is the AMSR-E snowmelt rate dramatically more gradual than the OL or the Gamma-DA simulations?

[Answer] We appreciate you for pointing this out. We carefully investigated the AMSR2 DA processes and found that there was a unit error when regridding the original AMSR2 SWE (10 km) to the OL grid (4 km) before the DA process. We originally thought the unit of AMSR2 SWE is "cm" so the SWE values were multiplied by 10 to make "mm" though the conversion was not needed as the original unit is "mm". This error caused the weird DA simulations with a later melt-out date than OL. We re-ran the AMSR2 DA simulations with the unit correction for all gamma lines and regenerated the time series (Figure 4) and boxplot (Figure 7). To help compare, the time series of the AMSR2 DA with/without the unit correction (SJ150) are provided below. The revised figures 4 & 7 are also attached.

Before the unit correction



After the unit correction



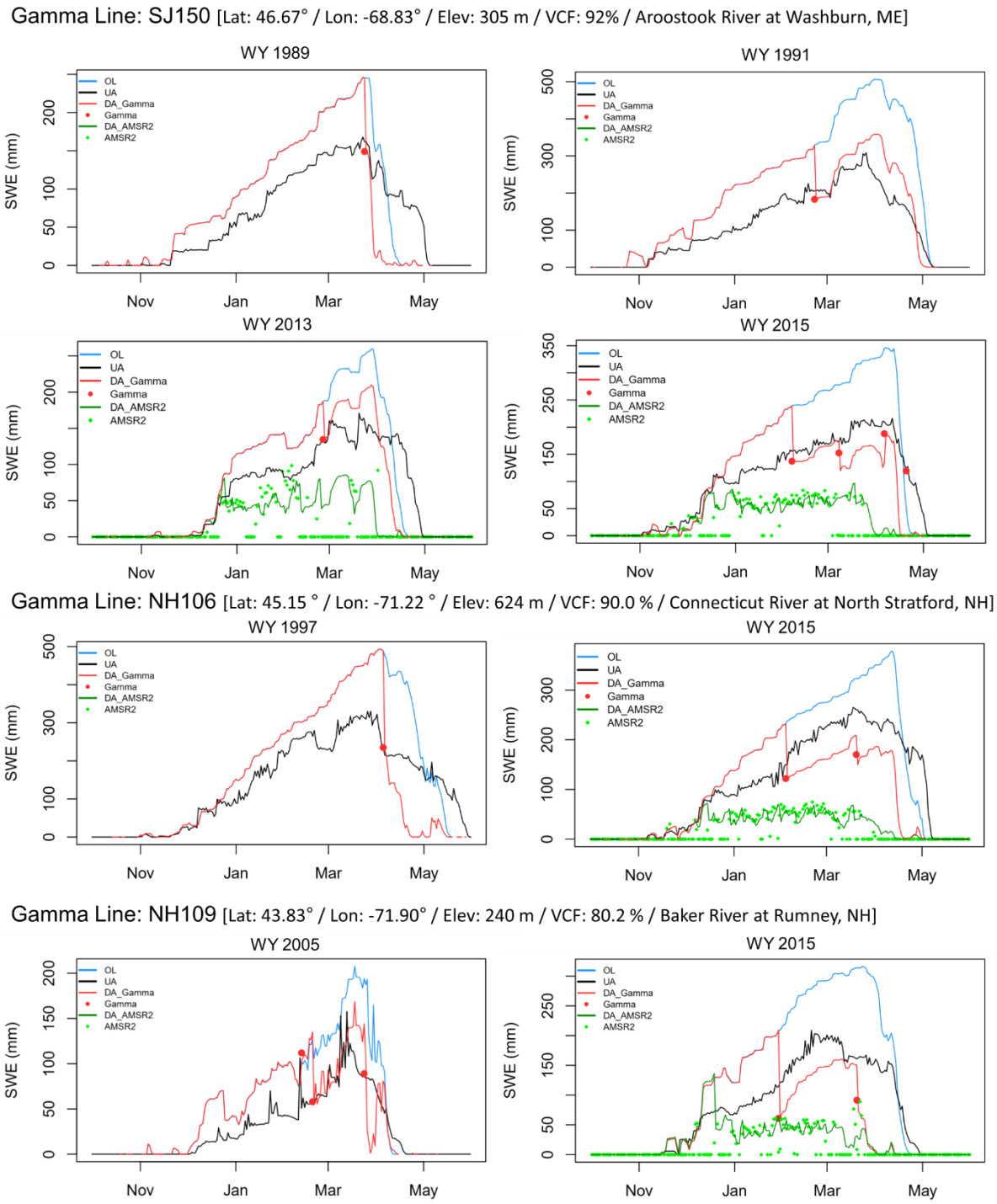


Figure 4. Examples of daily SWE time series of three gamma lines (SJ150, NH106, and NH109) with latitude (Lat), longitude (Lon), elevation (Elev), and vegetation cover fraction (VCF) for individual years including the open-loop (OL) and gamma data assimilated (DA_Gamma) Noah-MP SWE estimates along with the passive microwave SWE data from the Advanced Microwave Scanning Radiometer 2 (AMSR2) and AMSR2 data assimilated SWE (DA AMSR2).

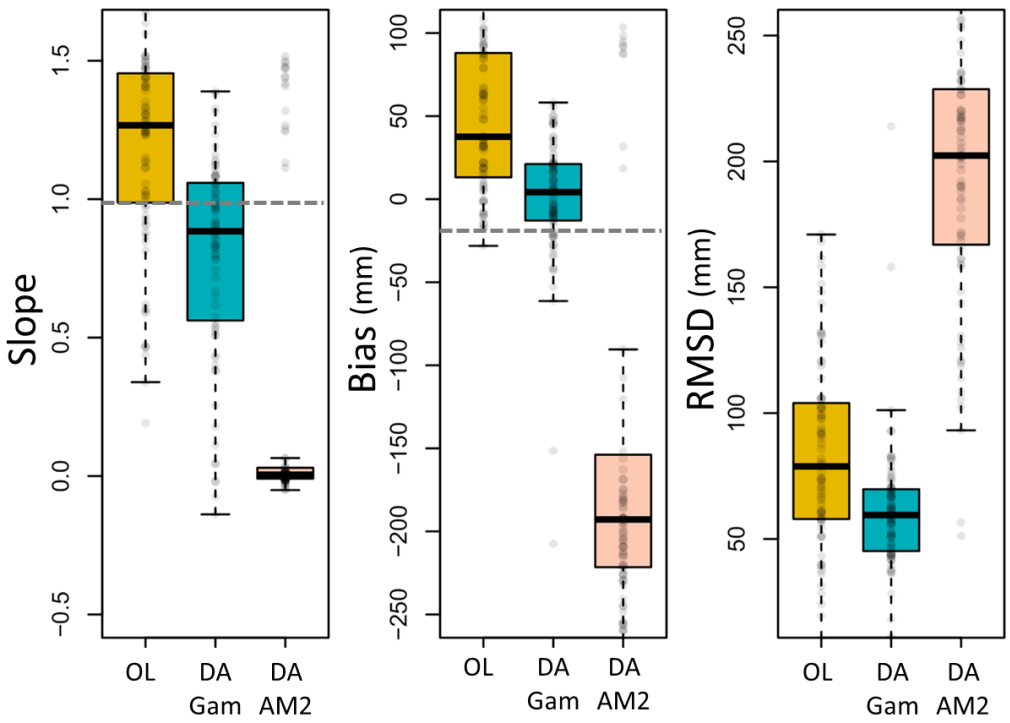
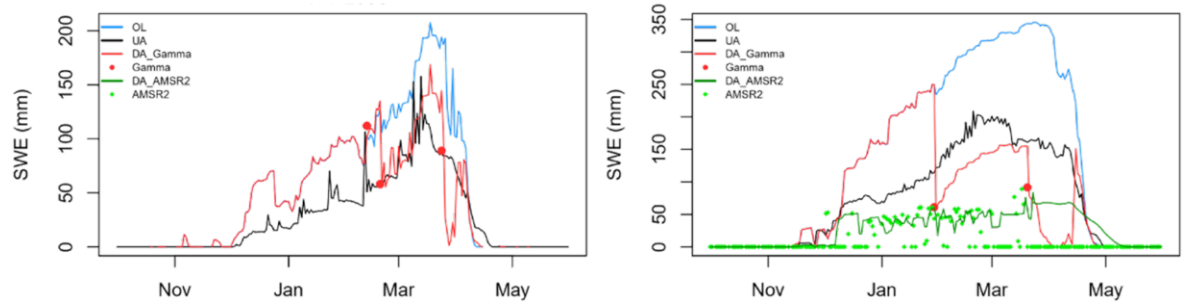


Figure 7. Comparison of the SWE estimation performance between the open-loop (OL), gamma DA, and AMSR2 DA as compared to the UA SWE at the 16km localization distance for the mutual DA effective accumulation periods.

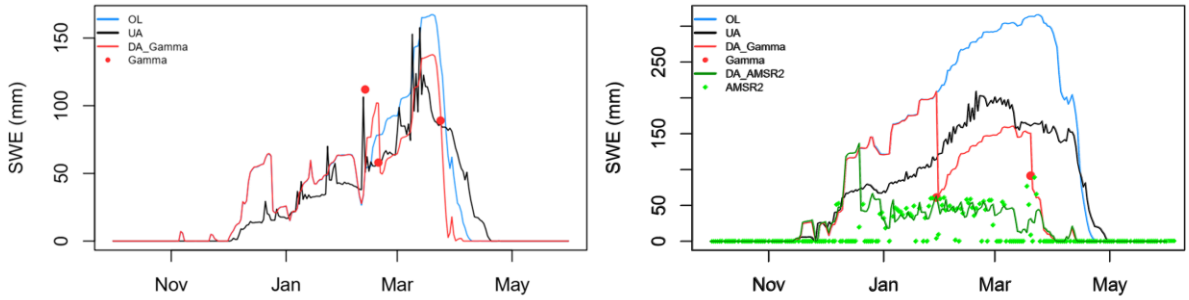
b. What is going on at the Rumney time-series that allows for a late-season spikes in SWE time-series in the Gamma-DA simulation that is not reflected in the OL simulation? My first assumption was that there was a late-season snow event each year, but then wouldn't that also show up in the OL simulation? I'm confused how there is this spike in SWE in the DA simulation that appears without a flightline gamma observation to explain it.

[Answer] We carefully investigated the time series at Rumney, NH, and found that there was an issue when calculating a grid-average SWE time series for the NH109 gamma line. A pixel that should be excluded was included in the calculation, resulting in the late-season spikes. Here is a time series plot with grid-average SWE values before/after the grid correction. The grid correction also led to slight changes in the OL time series. We replaced it with the new time series for NH109 in Figure 4.

Before the grid correction



After the grid correction



3. Additional model details would be helpful here in section 4. For example, if the model is run over a gridded domain, what is the grid-spacing? How does this grid-spacing compare to the MERRA forcing? Would it be helpful to show an outline of the model domain in figure 1? Is figure 1 already *showing* the model domain? Finally, while the authors indicated in their responses that the MERRA-2 data was interpolated to the LIS grid, was it *downscaled* to the terrain at all? (e.g., was temperature adjusted using a lapse-rate?). There is a lot of terrain in NH and ME that the flightlines sample, and the MERRA forcing almost certainly doesn't capture it adequately.

[Answer] Thanks for the suggestion. The model was run over a gridded study domain (LAT: 42.76 to 47.44 / LON: -72.29 to -67.85; Figure 1) with a grid spacing of 0.04° while the spatial resolution of the MERRA-2 forcing is 0.625° (latitude) × 0.5° (longitude). Therefore, the original MERRA-2 forcing was downscaled to the model grid using the bilinear method within LIS. During the interpolation, topographic correction methods were applied for air temperature, air pressure, humidity, and longwave radiation based on the lapse rate (Cosgrove et al., 2003), and for longwave radiation by considering the impact of topographic slope and aspect (Dingman, 2002). These explanations were included in the revised manuscript.

*** References**

Cosgrove, B.A., Lohmann, D., Mitchell, K.E., Houser, P.R., Wood, E.F., Schaake, J.C., Robock, A., Marshall, C., Sheffield, J., Duan, Q., Luo, L., Higgins, R.W., Pinker, R.T., Tarpley, J.D., Meng, J.: Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *Journal of Geophysical Research*, 108(D22), 8842. https://doi.org/10.1029/2002JD003118, 2003.

Dingman, S.L.: Physical Hydrology (2nd ed.), Prentice-Hall, 2002.

Minor comments:

Line 137: What does UA stand for, I think this is the first time this acronym shows up, please define it.

[Answer] We defined the University of Arizona (UA) here.

Line 139: What is the snow density parameterization, reference?

[Answer] We added a related reference, Dawson et al. (2017).

Dawson, N., Broxton, P., & Zeng, X.: A new snow density parameterization for land data assimilation. Journal of Hydrometeorology, 18(1), 197–207. https://doi.org/10.1175/JHM-D-16-0166.1, 2017.

Line 191: The BATS and CLASS albedo schemes are specifically for the snow albedo, not the total ground albedo, please correct this.

[Answer] Thank you for pointing this out. We corrected this.

Line 197: Would it be helpful to elaborate on the purpose of the ensemble here, e.g., to generate model uncertainty metrics for the DA?

[Answer] Thank you for the suggestion. We added the purpose of the ensemble here.

Then, using a restart file generated in the first step, an additional 3-year spin-up, from 1 January 1981 to 1 March 1984, was conducted using 20 ensemble members to generate model uncertainty metrics for the DA.

Line 224: "of limited used" should be "of limited use"

[Answer] We used "can be limited to be used"

Line 231: Consider replacing "the degree of the SWE updates" with "the magnitude of the SWE DA adjustment" or something along those lines.

[Answer] We replaced this with the suggested words.

Line 244 – 249: I suggest rewording "However, this was a consequence of the fact that the overestimated SWE during the accumulation season and early in the melt season was offset by the underestimated SWE during the snowmelt season (i.e., April and May). When the gamma SWE observations exist during the accumulation period (which is a typical case), DA corrected the overestimated SWE, whereas it further underestimated SWE in the snowmelt season (**Figures 3** and **4**), resulting in the increased (negative) bias, as presented in **Figure 2**."

As

However, this was a consequence of the fact that in correcting the overestimated SWE during the accumulation season, the DA introduced a greater underestimate during the

melt season (Figures 3 and 4).
[Answer] We reworded this as the Reviewer suggested. Thank you.

Line 251: I suspect that the r-value decrease is due almost entirely to the increase in the number of zero – to – not-zero comparisons involved in the analysis. The physical lower-boundary of snow as a variable really complicates a lot of these statistical metrics. Perhaps a BETTER comparison would be to only compare data-points where both the simulated and observed SWE are greater than zero.

[Answer] Thank you for the suggestion. We had actually considered those comparisons in the analysis. However, we concluded that the results based on zero-to-not-zero comparisons would be more appropriate because a calculation using non-zero values only could generate misinterpretation by removing the portion of adjusted SWE during the early snowmelt though this may result in a better R-value (we also think this would make sense for consistency with other statistical metrics such as bias).

Line 261: Again, I have trouble with the statement "limited model physics" here. Pending changes made to address major comment 1 above, I think this can be addressed with a simple change of wording to something along the lines of (changes in ***bold/italics***):

"However, the assimilation of the airborne gamma SWE measurements was not able to improve the snow ablation timing ***due to sparse gamma data during the spring in combination with the overall poor model performance during the melt season***" Essentially, any rewording that more generally acknowledges the poor performance in this specific instance, rather than speculating that there is something intrinsically wrong with the model physics.

[Answer] We agreed with the point. We changed the original statement with the suggested one the Reviewer made.

Line 264: "single gamma SWE" should be "single gamma SWE flight?"

[Answer] Changed.

Line 273: I suggest that you remove "In the figure" from the beginning of the sentence.

[Answer] Agreed. We removed.

Line 275: "has low bias and RMSD than OL SWE". This is an incomplete though, the model has a low bias, and a ??? RMSD compared to OL SWE? Higher? Lower? Equivalent? Please fix.

[Answer] We fixed by adding "lower RMSD".

Line 275: Suggested replace: "DA performances show relatively" to "DA led to"

[Answer] Replaced.

Line 279 (and more generally throughout the manuscript): Consider replacing the word "updated" with "improved." Using the word updated is ambiguous regarding whether or not the

simulation was improved. I'll note it here, but I recall seeing the use of "updated" where a better word choice is possible at other places throughout the manuscript as well.

[Answer] Thank you for your suggestion. We agreed and revised the word "updated" with "improved" throughout the manuscript. In certain statements whether or not the DA simulation was improved, we remain the word.

Line 294: consider replacing "surrounding areas, where gamma flights do not exist" with "In areas surrounding the gamma-flights"

[Answer] Replaced. Thank you.

Figure 6 and Lines: 297 – 307: This is a nice result. Is there a way to assess statistical significance in this comparison? E.g., at what localization distance are the improvements compared to the OL no longer statistically significant (or are all of these significant)? The results here are somewhat convincing visually, so I leave any decision to include a statistical significance analysis here to the authors.

[Answer] We agreed that Figure 6 here is sufficient enough to determine the magnitude of the improvements of the DA with different localization distances. We remain the current version with no inclusion of the significant test results.

Line 297: what is the "effective surrounding areas"? is this a fixed number for ALL localization radii? Or only gridcells within the localization radii?

[Answer] This means grid cells within the localization radii. We edited this statement as below.

"The OL/DA statistics in the figure are calculated using domain-averaged time series of OL/DA SWE for grid cells within a given localization distance with the corresponding UA SWE."

Line 301: Why is the OL RMSD decreasing as localization radii?

[Answer] This is because the spatial variability of SWE was smoothed and the domain-average values became similar, the domain-average OL and UA SWE for larger localization radii tend to have a lower RMSD.

Line 316: replace "reduced the model SWE errors" with "improved model SWE"

[Answer] Done.

Line 317: There is an errant open parenthesis here.

[Answer] We removed this.

Line 332: replace the word "ground" with the word "snow"

[Answer] Done.

Line 334: add, "snow albedo" after "BATS and CLASS"

[Answer] Added.

Line 340: replace “but it still has” with “but did not improve”

[Answer] Done.

Line 340: replace “combinational” with “combined”

[Answer] Done.

Line 341: “fully-implicit” should be “the fully-implicit”

[Answer] We edited it.

Line 344: replace “worth to note” with “worth noting”

[Answer] Replaced. Thank you for the detailed corrections.